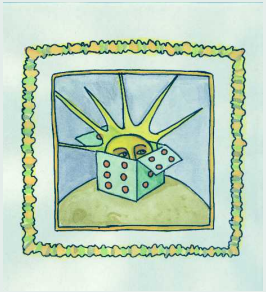
Quantifying network performance based on the ozone standard

Douglas Nychka and Eric Gilleland
Geophysical Statistics Project,
National Center for Atmospheric Research
www.cgd.ucar.edu/stats

- The problem: A nonlinear, seasonal statistic
- A useful space/time model for 8 hour
- Model for RTP,NC
- Thinning the network



National Center for Atmospheric Research

$\approx$ 1000 people total, several hundred PH D (physical) scientists,
half the budget ($\approx$ 60M) is a single grant from NSF-ATM

Research on nearly every aspect related to the atmosphere

Climate, Weather, the Sun, Ocean/atmosphere, Ecosystems, Economic impacts, Air quality, Instrumentation, Scientific computing and ...

Statistical methods for the geosciences

The problem

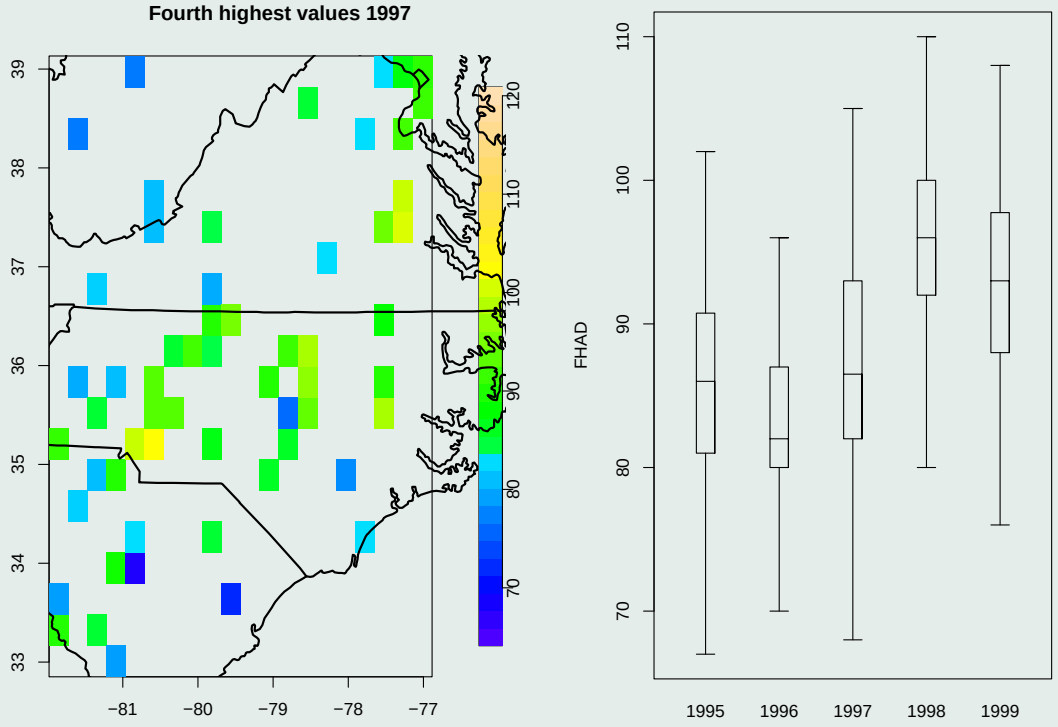
A suggested ozone pollutant standard is based on the fourth highest (max) 8-hour daily average recorded during the year. Compliance is related to a three year average being less than 85 PPB.

Although it is straight forward to build spatial statistical models for the daily ozone field, the extension to the fourth highest measurement is difficult ....

- Time dependence
- Nongaussian statistic (extreme value)
- Covariance structure

An example for North Carolina

Fourth highest daily average (FHDA) values.



Approach

Conditional Simulation

For unmonitored locations find the conditional distribution of the FHDA. The distribution of the fields does not have a closed form and so we just generate samples from it.

Space-time model

In order to simulate from the FHDA field one needs a model for the temporal and spatial dependence of daily ozone. Transformed and scaled daily ozone follows an autoregressive model with spatially correlated shocks.

Model components

Transformation:: $O(\mathbf{x}, t) =$ 8-hour ozone at location $\mathbf{x}$ and time t.

$$u(\mathbf{x}, t) = \frac{O(\mathbf{x}, t) - \mu(\mathbf{x}, t)}{\sigma(\mathbf{x})}$$

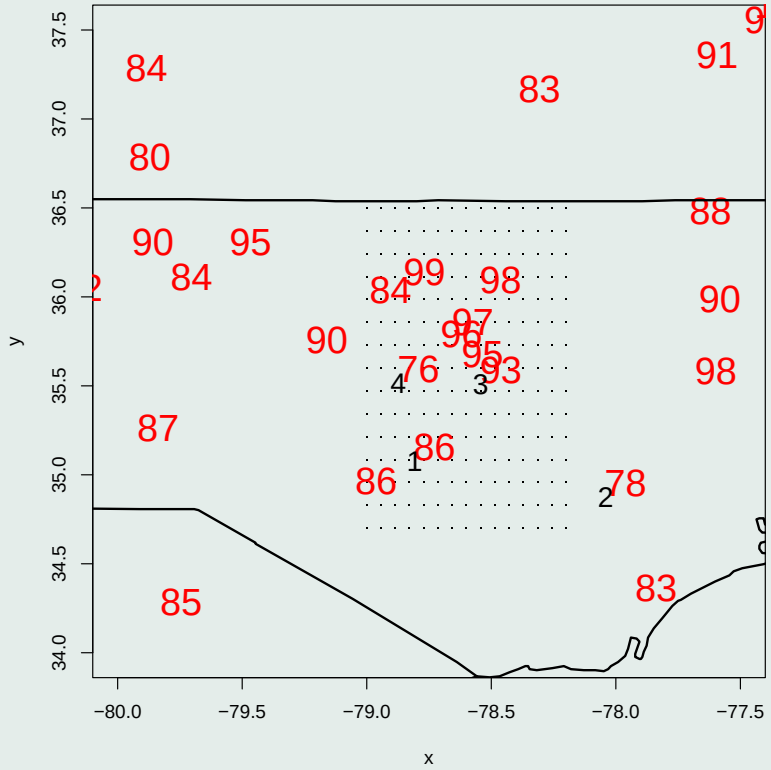
Autoregression: $u(\mathbf{x}, t) = \rho(\mathbf{x})u(\mathbf{x}, t) + e(\mathbf{x}, t)$

Spatial dependence: $e(\mathbf{x}, t)$ uncorrelated over time and stationary over time.

$$COV(e(\mathbf{x}, t), e(\mathbf{x}', t)) = (1 - \rho(\mathbf{x})^2)k(||\mathbf{x} - \mathbf{x}'||)$$

Under the assumption of multivariate normality one can generate fields of daily ozone conditional on the observed values.

Test case domain

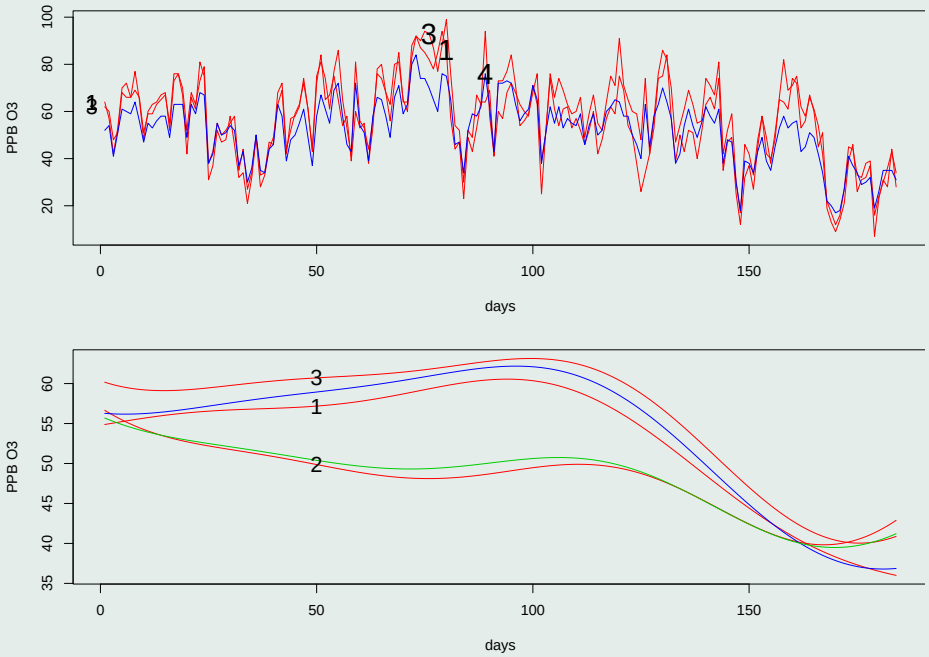


Transformation

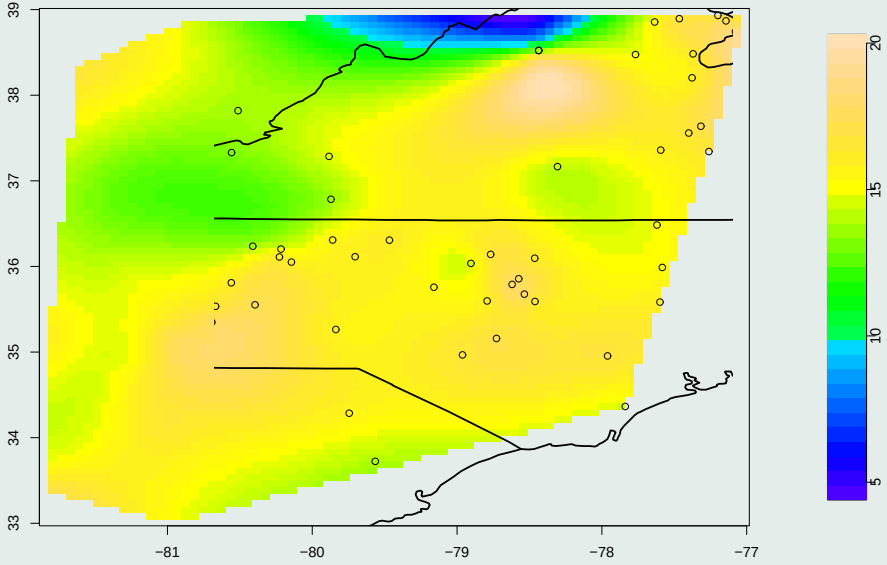
$\mu(\mathbf{x}, t)$ was estimated for each station location using OLS on a sine/cosine expansion and then smoothed over space using PC applied to regression coefficients. Extrapolation to unmonitored locations using interpolating thin plate splines (TPS) .

Example of time series

Data and estimated seasonal cycle.

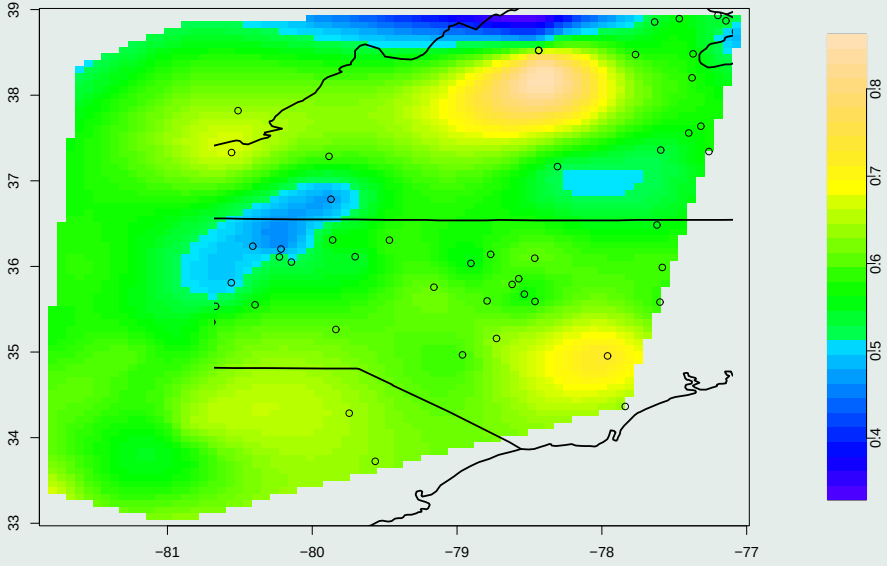


$\sigma(\mathbf{x})$ also based on TPS interpolation of station estimates.



Autoregressive model

$\rho(\mathbf{x})$ found from autoregression on transformed station data and then extrapolated using TPS

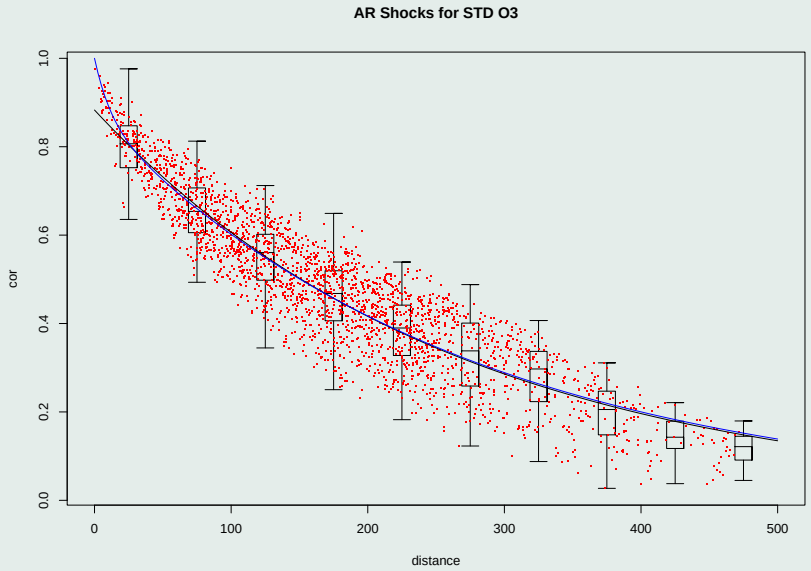


Spatial dependence

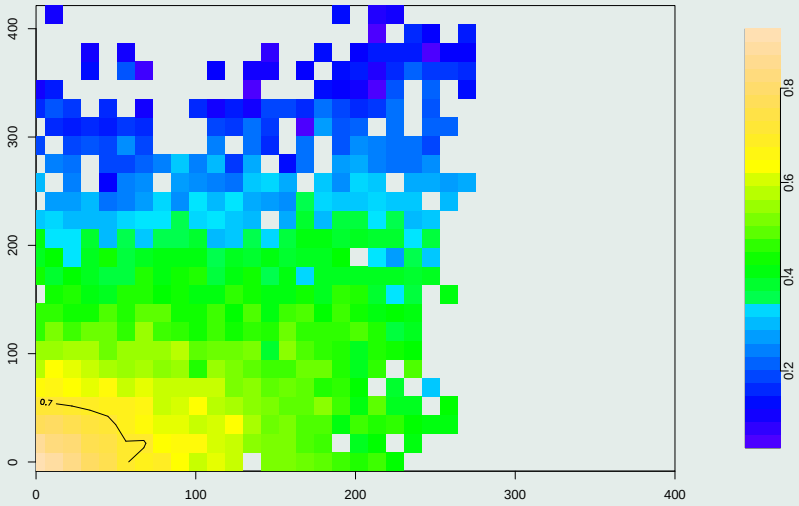
Correlogram of shocks suggests a mixture of exponential covariances

$$k(d) = \alpha e^{-d/\theta_1} + (1-\alpha)e^{-d/\theta_2}$$

with $\alpha = .09$, $\theta_1 = 18$ (miles) and $\theta_2 = 270$ (miles)



Anisotropy?



The algorithm ...

First discretize this problem
$\mathbf{o}_t$ daily ozone values on a grid and including the stations locations.

$$\mathbf{o} = \begin{pmatrix} \mathbf{o}^s \\ \mathbf{o}^g \end{pmatrix} \tag{1}$$

Generating one year

Start with initial field: $\mathbf{o}_0$

1. *spatial shock* sample from $[\mathbf{e}_t^g | e_t^s]$
2. *propagate* $\mathbf{u}_t = \rho \mathbf{u}_{t-1} +$ conditional shocks
3. *back transform* $\mathbf{O}_t = \mathbf{u}_t \sigma + \mu$

Repeat for entire season.
For each location compute FHDA.

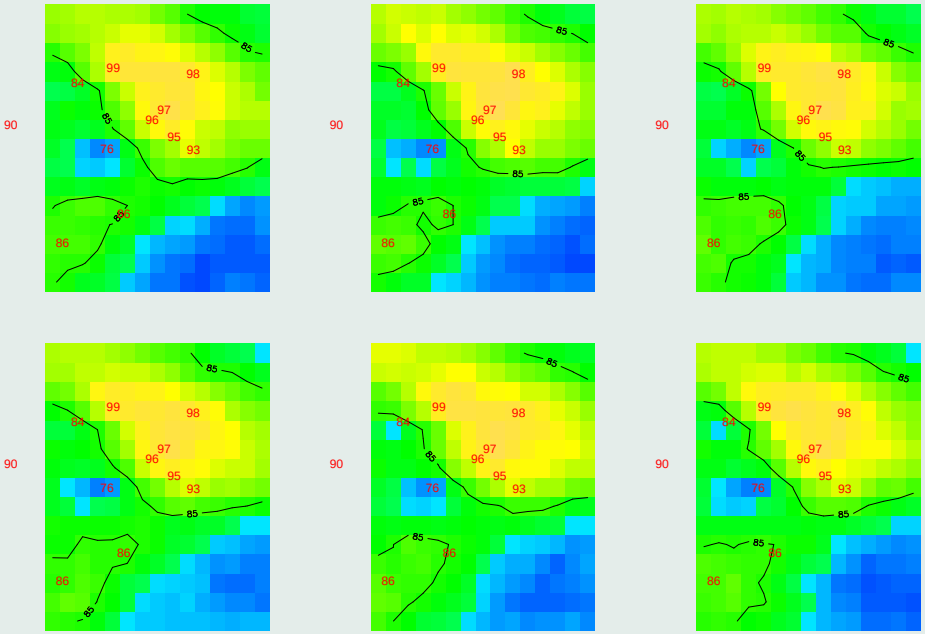
Corner of box, two samples from conditional dist.

Top: Mean (blue) samples (red)

Below: Differences from mean

Inference/Posterior

Repeat simulations of year to accumulate a distribution of FHDA



The main subtlety

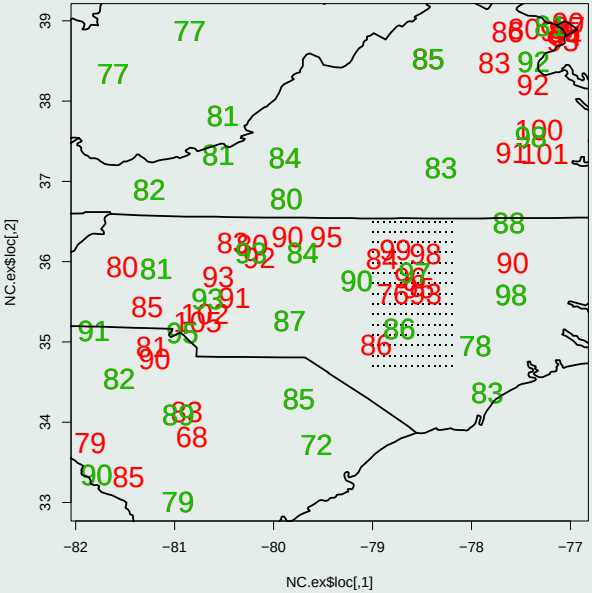
Formally it is important not to reverse the orders of computing FHDA and finding the mean.

In step 2 : *spatial shock* is a sample from $[\mathbf{e}_t^g | e_t^s]$

One generates random samples in instead of finding conditional mean.
Averaging happens at the very end of the process
e.g. in finding a posterior mean for the FHDA at an unmonitored location.

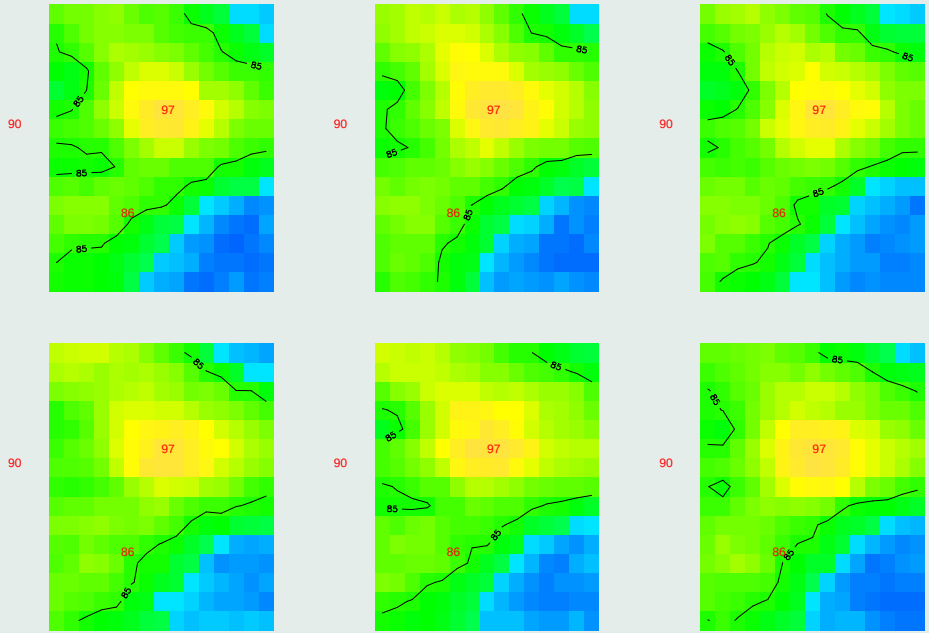
Sensitivity to network thinning

Suppose the 72 stations is thinned to 36 (green) using a geometric space-filling criterion.

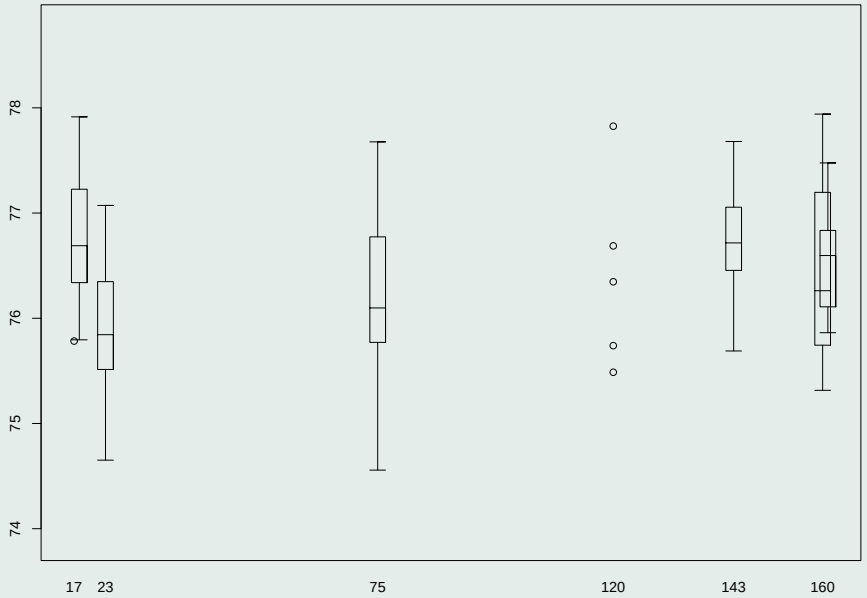


How does the accuracy in determining the FHDA change?

Conditional distribution

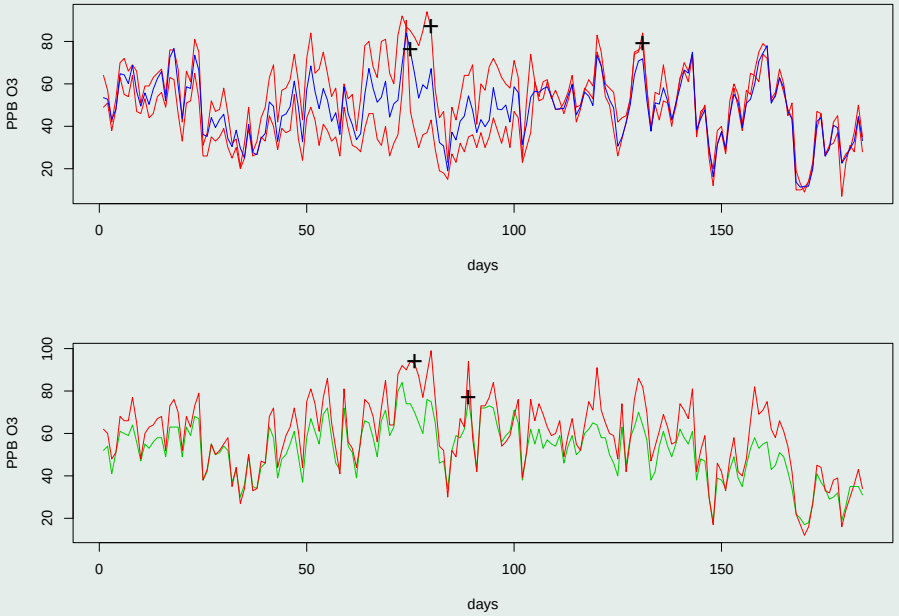


Results for corner



Strangeness

Low value from station 34 seem strange. Due to small variance for this year.



Shortcuts

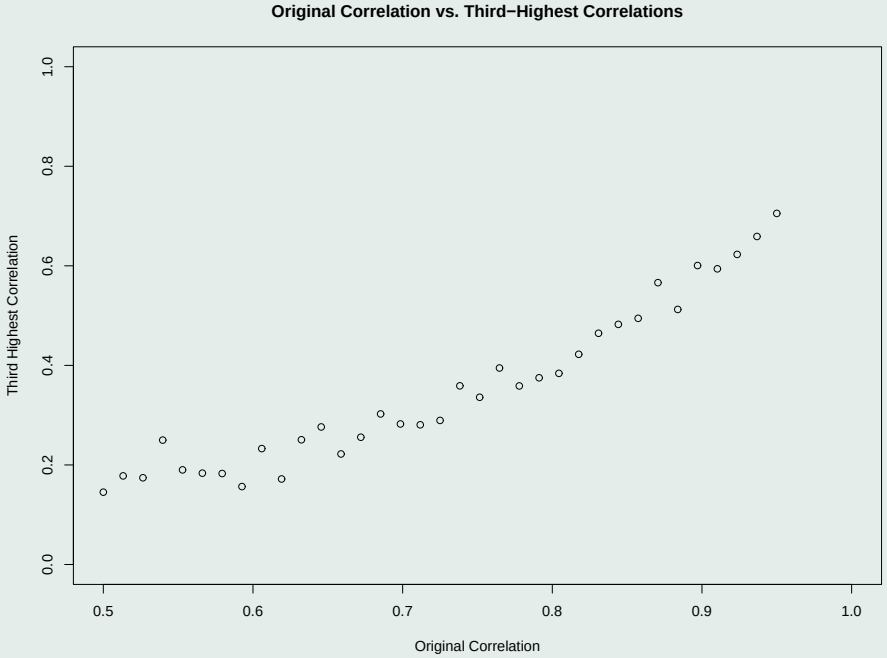
This seems like a lot of work just because we don't know the covariance!
Especially to implement analysis of FHDA in an interactive framework.

Bivariate extremes problem

Suppose that $(X_k, Y_k), 1 \leq k \leq n$ distributed bivariate normal with correlation, ρ.
What is the joint distribution of FHDA? Is there any hope of being simple?

Estimated correlations for FHDA

At least the bivariate distribution appears Gaussian.
Relationship among correlations



Discussion

- Simple space-time models can be estimated and sampled
- Results are sensitive to modeling tails.
- FHDA may be close to multivariate normal under normal assumptions
- Need to build more uncertainty/varaiblity in the model.